基于《知网》的多种类型文献混合自动分类研究*

李湘东[1,2]　刘　康[1]　丁　从[1]　高　凡[1]
[1](武汉大学信息管理学院　武汉　430072)
[2](武汉大学信息资源研究中心　武汉　430072)

摘要: 【目的】解决由于不同类型文献而产生的特征不匹配等问题，提高待分类文本的分类效果。【方法】使用与待分类文本属于不同文献类型的文本作为语料库的训练集，引入第三方资源《知网》进行语义特征扩展。【结果】利用该方法在网页、图书、非学术性期刊、学术性期刊 4 种类型文献上进行分类实验，与未经过扩展的分类方法相比，分类准确率提高 1.2%至 11.0%。【局限】未对每一种文献类型都使用公开语料进行测试，因此本文方法的通用性和实验结果的客观性有待进一步检验。【结论】实验结果表明，该方法具有一定的可行性和实用性，在不同程度上可以消除不同类型文献之间的语义差异，从语料库构建和特征扩展两个途径提高文本自动分类的分类效果。

关键词: 第三方资源　知网　特征扩展　语义差异
分类号: TP393　G35

1　引　言

随着互联网的迅猛发展，网络上的信息资源日益剧增，人们可以从互联网上源源不断地获取各种形式的信息，如文本、图片、音频、视频等。而文本可以来自于网页、图书、学术期刊论文等众多文献类型，人们可以获得同一主题下具有不同内涵、质量、发布速度的信息。因此，利用文本分类技术将这些文本信息分门别类，以便它们能够更加快捷、有效地被分类组织或检索的研究，具有较高的应用和实用价值。

文本自动分类涉及训练集构建、特征选择、分类算法等众多环节。自动分类研究中训练集与待分类文本通常使用同一类型的文献，但信息资源管理领域的相关研究表明，使用与待分类文本属于不同类型的文献构建训练集时，也有可能提高待分类文本的分类效果[1-3]，因此不同类型文献之间的混合分类成为提高分类效果的途径之一。但是，这些研究没有注意到不同类型文献之间在用词习惯、写作风格上具有不同特点，使得来自训练集的特征和待分类文本的特征，原本是表达同一概念的，却出现不能很好地进行匹配的问题。因此，使用与待分类文本属于不同类型的文献作为训练集开展自动分类时，需要进一步研究适当的方法克服训练集与待分类文本之间在用词及语义等方面的差异，以提高两者之间共同特征的数量，进一步改善分类性能。

本研究借助第三方资源对属于不同类型文献的训练集和待分类文本进行特征扩展，通过扩展特征的数量和语义，使训练集和待分类文本中表达同一概念的特征之间增加匹配的可能，以达到由不同类型文献所构成的训练集和待分类文本之间具有更多的共同特征的目的。本文实验数据包括经过科学分类、且长期积累的图书或期刊论文等学术性文献，以及网页、时事

通讯作者：李湘东，ORCID：0000-0001-9031-8482，E-mail：xli_xiao@hotmail.com。
*本文系国家社会科学基金项目“多种类型文本数字资源自动分类研究”(项目编号:15BTQ066)的研究成果之一。

周刊等新闻性、非学术性、更新频度较强的文本。按照文献类型的不同分别作为训练集和待分类文本(测试集)开展自动分类研究，提出一种基于《知网》[4]的语义特征扩展方法，并通过实验证明其能提高多种类型文献的分类效果。

2 研究现状和意义

2.1 国内外研究现状及发展动态

基于机器学习的文本自动分类研究，需要分类算法对训练集进行学习，并将学到的知识用于对测试集的分类。传统机器学习分类中，训练集和测试集通常使用同一类型文献，而多种类型文献混合分类可以利用已有或较易获取的训练集，对不同文献类型的测试集进行分类。其依据来源于迁移学习中的跨领域分类思想[5]。跨领域分类是国内外机器学习领域研究的前沿主题之一，其基本出发点是针对来自不同领域的训练集和测试集进行分类。所谓不同领域，是指训练集和测试集可以是不同的学科主题内容，也可以是不同的产品评论，训练集和测试集甚至可以分别使用不同语言的文本。文献[6-7]利用维基百科等第三方资源作为中介，对训练集和属于不同主题范围的测试集之间的特征进行关联，减少训练集与测试集之间因为主题内容不同而在语义特征上的差异，其目的就是为了在训练集与测试集之间构造具有更多共同特征的特征空间，从而提升分类效果。本文利用这种跨领域分类的思想，将《知网》作为第三方资源，用于增加不同文献类型的训练集和测试集之间特征匹配的可能性，是一种跨文献类型分类或跨源文献分类问题。

短文本特征扩展也是近几年文本自动分类领域研究的热点问题之一，其核心思想是通过特征扩展扩大训练集与测试集之间共同的特征数量或语义信息，从而提高分类效果。例如，文献[8]以维基百科词语相关概念集合作为特征扩展词集，利用维基百科中概念之间的链接、类别关系分别对训练集和测试集的短文本进行特征扩展，通过扩展特征数量提升分类性能；又如，文献[9]从丰富特征词语义的角度出发，抽取领域高频词作为特征词，基于《知网》从语义方面将训练集和测试集中的特征词扩展为概念和义元，并利用不同概念所包含相同义元的信息量计算特征词的相似度实现分类，也提升了分类效果。

本研究将短文本分类研究中特征扩展的方法应用于不同类型文献所构成的训练集和测试集上，通过扩展特征的数量和语义，使不同类型文献之间具有更多的共同特征，从而帮助提高分类效果。

2.2 研究意义

机器学习的文本自动分类是自动分类的主流方式，其基本过程主要包括：构建语料库、文本建模、特征选择、特征扩展、选择并实现分类算法等环节。在人工智能领域，自动分类的主要研究内容是除构建语料库以外的其他环节。而在信息管理领域，由于文献是其主要研究和应用对象，关于文献的分类、内容特征(如主题)、类型以及特性，有众多的研究成果。因此，在开展自动分类研究时，很自然地就比较重视语料库环节中训练集和测试集作为文献的文本特性，并试图将其应用到提高自动分类效率上。本文借鉴信息管理领域中，单纯使用不同类型文献分别作为训练集和测试集以提高分类性能的研究成果，从缩小训练集和测试集之间因文献类型的不同而可能产生语义差异的角度，试图进一步从语料库出发提高分类效果。

3 研究方法

3.1 基于《知网》语义特征扩展的分类框架

为解决训练集和测试集因为不同类型文献之间的差异而产生特征上的不匹配，本文按照文献[10]中有关特征扩展的基本思想，提出一种针对不同类型文献开展特征扩展的文本分类方法。具体分类框架如图 1 所示：

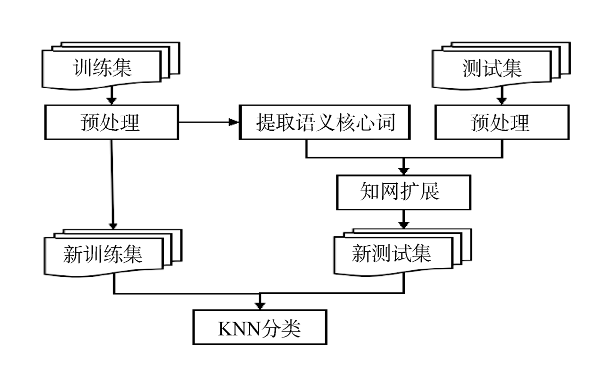


图 1 基于《知网》语义特征扩展的文本分类框架

(1) 预处理。分别对属于不同文献类型的训练集和测试集文本进行分词、去停用词等预处理，预处理之后可以得到每篇文章对应的初始特征集合。

(2) 计算训练集文本中特征词的 TF-IDF 权重，提取大于某一阈值的特征词构成语义核心词集。

(3) 对经过预处理的待分类文本 d，借助《知网》的语义词典计算 d 中每个特征词与训练集语义核心词集中各特征词之间的语义相似度，将相似度值大于某一阈值的特征词扩展到文本 d 中，获得扩展后的待分类文本。使得测试集中具有相近语义的特征，能够通过《知网》得以扩展、与训练集中的特征匹配。

(4) 采用 KNN 算法，计算待分类文本与训练集中语义核心词集的相似度，将相似度最高的类别分配给该待分类文本。

3.2　训练集语义核心词集的获取

TF-IDF 权重被广泛应用于文本分类进行特征权重计算，其主要思想是如果某个特征项在一个文本中出现的词频高，并且在其他文本中很少出现，则认为此特征项具有很好的类别区分能力[11]。因此，本文将训练集中 TF-IDF 权值较高的特征词作为语义核心词进行特征扩展，统计训练集中每个特征词在每个类别中的 TF-IDF 值，取大于阈值的词作为语义核心词。具体流程如下：

输入：训练集 D，特征词 TF-IDF 阈值 weight
输出：训练集语义核心词

①对训练集进行词性过滤，仅保留对分类影响较大的名词、动词和形容词；

②计算每个特征词在各文本中的 TF-IDF 权值；

③对每一篇文本中的特征词进行归一化处理，设特征词 i 在文本中的 TF-IDF 值为 w_i，按照公式(1)进行归一化处理：

$$w_i = \frac{w_i}{\sqrt{\sum_{i=1}^{n} w_i^2}} \qquad (1)$$

④取在每类中占比大于阈值 weight 的特征词作为训练集的高频词。

3.3　基于《知网》的语义相似度计算

本文借助《知网》计算特征词之间的语义相似度，以此挖掘由不同文献类型文本所构成的训练集和测试集之间的相关关系，由此可见，基于《知网》的语义相似度计算是本文特征扩展方法的基础。在《知网》的结构中，词语由义项表示，即一个词语可以表示为多个义项，而一个义项又由义原来表示。因此，词语之间的相似度可以通过计算其义原间的相似度得到。

(1) 义原相似度计算

《知网》中的义原树是由义原之间的上下位关系构造成一个树状结构的层次体系，本文通过计算义原树中义原间的最短路径距离计算义原相似度。设两个义原的最短路径距离为 d，则这两个义原之间的相似度计算公式如下[12]：

$$\mathrm{sim}(p_1,p_2) = \alpha/(d+\alpha) \qquad (2)$$

其中，p_1,p_2 表示两个义原，d 为 p_1 和 p_2 在义原树中的最短路径距离，α 是可调参数。

文献[13]认为仅考虑义原之间的最短路径距离不能准确地计算两者的相似度，提出考虑义原层次深度的义原相似度计算方法。其主要思想是：对于相同最短路径距离的两个义原，层次越深，义原描述的含义越具体，应该赋予更大的相似度权重。计算公式如下：

$$\mathrm{sim}(p_1,p_2) = \frac{\alpha \times \min(depth_{p_1}, depth_{p2})}{\alpha \times \min(depth_{p_1}, depth_{p2}) + d} \qquad (3)$$

其中，$depth_{p_1}, depth_{p2}$ 分别为义原 p_1 和 p_2 在义原树中的深度，d 为 p_1 和 p_2 在义原树中的最短路径距离，α 是可调参数。本文采用公式(3)计算义原间相似度。

(2) 义项相似度计算

在《知网》中对义项的描述一般分为 4 类[14]，即主要特征、次要特征、关系义原特征以及关系符号特征描述式。义项的相似度为语义描述式的各个对应组成部分间的相似度的加权和。义项 s_1 和义项 s_2 的相似度计算如下[15]：

$$\mathrm{sim}(s_1,s_2) = \sum_{i=1}^{4} \beta_i \times \mathrm{sim}_i(s_1,s_2) \qquad (4)$$

其中，$\mathrm{sim}_i(s_1,s_2)$ 分别为义项 s_1 与 s_2 的语义描述式中的各组成部分之间的相似度，β_i 为各部分的相似度比重，$\beta_1+\beta_2+\beta_3+\beta_4=1$，且 $\beta_1 \geqslant \beta_2 \geqslant \beta_3 \geqslant \beta_4$。文献[12]给出了义项的语义描述式中各个组成部分的相似度计算方法，义项的相似度计算最终可以转化为义原间的相似度计算。

(3) 词语相似度计算

文献[15]认为两个词语之间的相似度就是它们在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性大小。假设有两个词语 w_1 和 w_2，若 w_1 含有 m 个义项：$s_1=\{s_{11},s_{12},\cdots,s_{1m}\}$，$w_2$ 含有 n 个义项：$s_2=\{s_{21},s_{22},\cdots,s_{2n}\}$，则词语 w_1 和 w_2 的相似度

为各个义项所有组合中相似度的最大值，采用了最大匹配的方法。计算公式如下[15]：

$$\mathrm{sim}(s_1,s_2)=\max_{i,j}\mathrm{sim}(s_{1i},s_{2j}) \qquad (5)$$

4 实验设计与分析

4.1 实验材料

本研究分别从搜狗语料库[16]、馆藏目录、电子期刊数据库等信息资源中获取网页、图书和期刊等三种类型的文献，其中，期刊进一步分为学术性期刊和非学术性期刊。网页文献选取搜狗语料库中体育、IT 和军事三个类别的文本构成实验材料。图书文献取自某大学图书馆的馆藏目录 OPAC，选取中国图书分类法分类体系下体育、计算机技术和军事三大类中部分图书的书目信息，提取其中的书名和摘要等内容构成实验材料的文本。期刊按照中国图书分类法，选取 CNKI 中体育、计算机技术和军事三大类的部分期刊。

本文对以上 4 种类型文献各建立多套实验材料并重复开展实验。每套实验材料包括一种类型文献的训练集和测试集，均由体育、计算机技术和军事三个大类构成，每一个类型文献分别由 600 篇文本构成，共 2 400 篇。

4.2 实验方法与测评方法

不同的分类算法会对分类结果产生较为明显的影响，本实验选择经典的 KNN 分类算法构造分类器。此外，从理论上讲，Naive Bayes 分类算法也能够达到不错的分类效果，但其特征项独立假设并不严格成立，所以经常被用作其他方法的比较标准。本文将 KNN 算法与 Naive Bayes 算法进行比较，说明 KNN 算法的有效性。采用 KNN 分类算法将多种文献单独自动分类、混合自动分类与本文提出的基于《知网》语义特征扩展的混合分类方法进行对照实验，得出结论并进行验证。

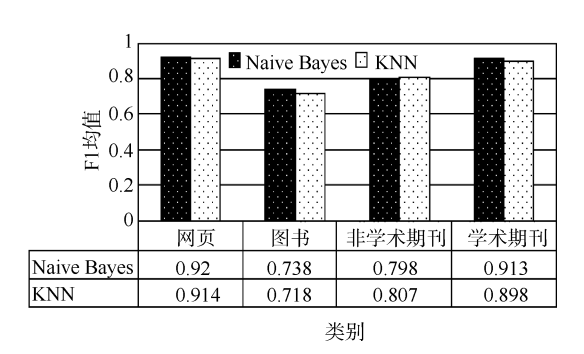
KNN 算法中的 k 值选取问题，本文使用一种自适应算法可以自动选取 k 值[17]，并且结果更为准确。训练集选取语义核心词集阈值根据预备实验结果测得取值为 0.8 时效果最佳。利用《知网》进行语义特征扩展中的各项参数根据经验[12]取值分别为 $\alpha=1.6$, $\beta_1=0.5$, $\beta_2=0.2$, $\beta_3=0.17$, $\beta_4=0.13$。分类效果的评价采用文本自动分类研究中通用的宏平均值 F1，它是对分准率和分全率的综合评价、代表分类系统的整体分类效果。

为消除不平衡数据对实验结果的影响，本文所有实验语料均采用平衡数据，包括各个类别包含大致相同数目的文本以及文本之间的长度差别不大，且训练集与测试集无重复。本文采取五折交叉验证法进行训练和分类，最后取 F1 的平均值作为实验结果[18]。

4.3 实验结果

(1) 各类型文献单独分类实验结果

各类型文献单独分类实验，是指训练集和测试集均为同种文献类型的文本时，对 4 种类型文献的实验材料分别开展五折交叉分类实验，保证训练集与测试集之间没有重复文本。分类效果如图 2 所示：



	网页	图书	非学术期刊	学术期刊
Naive Bayes	0.92	0.738	0.798	0.913
KNN	0.914	0.718	0.807	0.898

图 2 不同分类算法下各类型文献单独分类实验结果

使用 Naive Bayes 算法的实验结论与使用 KNN 算法的效果几乎一致，从而验证了 KNN 算法的有效性，因此选取 KNN 算法作为本文的分类算法。另外，由图 2 显示的实验结果可知，各类型文献在进行单独分类实验的时候都能取得不错的分类效果，分类准确率都在 70%以上。

(2) 各类型文献混合分类实验结果

各类型文献混合分类实验，是指训练集和测试集为不同文献类型的文本时，对 4 种类型文献的实验材料分别开展五折交叉分类实验。分类效果如图 3 所示。多种类型文献混合自动分类实验结果表明，4 种文献类型中，网页与非学术期刊之间的分类效果较好，均在 80%以上，并且其中以网页为训练集，非学术期刊为测试集的交叉分类效果达到了 83.9%，甚至好于非学术期刊的单独分类效果。而图书与学术期刊之间的分类效果较好，均在 70%以上。同样，在以学术期刊作为训练集，图书作为测试集的实验组里，分类准确率为 78.4%，也

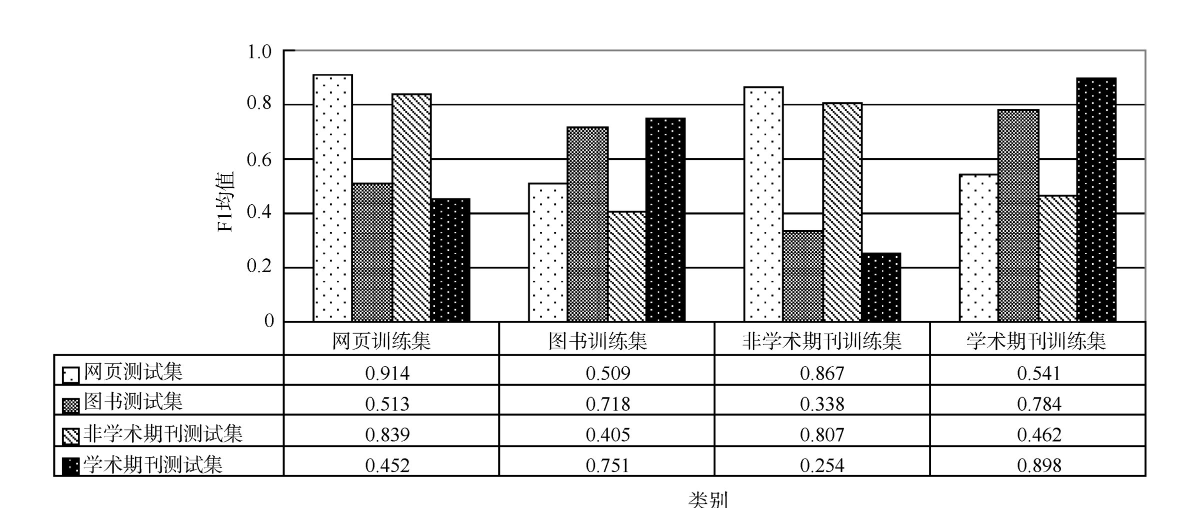


图 3　KNN 算法下各类型文献混合分类实验结果

高于单独用图书进行的分类实验准确率 71.8%。这证明了不同类型文献之间进行混合分类的合理性。

另外，4 种文献类型中，网页与图书、学术期刊，图书与网页、非学术期刊之间的分类效果较差，均在 60%以下。由此可知，不同文献类型的训练集和测试集的选择对分类效果的影响也是显著的。相互匹配的文献类型之间甚至可以获得比单独文献分类还要好的分类效果，而不匹配的文献类型之间往往难以取得较高的分类效果。这说明不同文献类型的组合对分类效果的影响是非常显著的。

(3) 基于《知网》语义特征扩展的混合分类实验结果

基于《知网》语义特征扩展的混合分类实验，是指将 4 种文献类型的文本分别作为训练集和测试集(包括训练集与测试集使用同一文献类型)，使用本文提出的特征扩展方法，将《知网》作为第三方资源对测试集进行特征扩展，再进行同样的五折交叉分类实验。分类效果如图 4 所示：

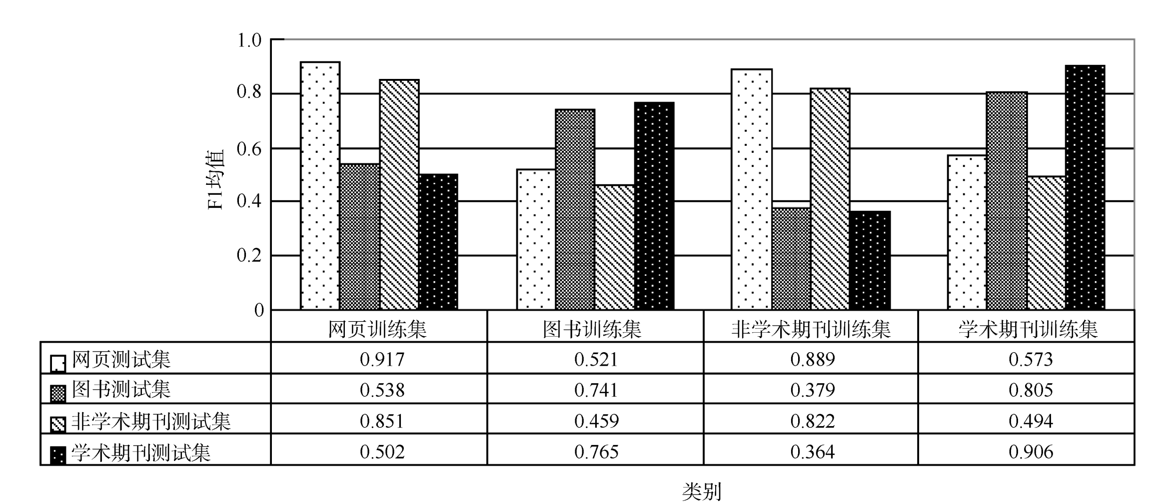


图 4　基于《知网》语义特征扩展的混合分类实验结果

基于《知网》语义特征扩展的混合分类实验结果表明，经过《知网》的语义特征扩展之后，4 种文献类型的单独分类效果有不同程度的提升，如以图书作为训练集和测试集的分类效果从 71.8%上升到 74.1%，原来分类效果较好的以网页作为训练集和测试集的分类效果从 91.4%上升到 91.7%；而 4 种文献类型的交叉分类效果也有较为明显的提升。其中，匹配程度高的文献增加的幅度较小，如非学术期刊为训练集，网页为测试集的分类准确率从 86.7%上升到 88.9%。而原本匹配程度不高的文献效果提升明显，如非学术期刊

为训练集，学术期刊为测试集的分类准确率从 25.4% 上升到了 36.4%。这说明原本匹配度低的文献之间具有更大的提升空间。

4.4 实验结果分析

实验(1)分别比较 KNN 与 Naive Bayes 算法对各文献类型文本进行单独分类实验，经比较两种算法效果几乎一致，本文选取 KNN 算法进行后续实验。实验(2)将不同文献类型文本进行混合自动分类，由图 3 中结果可知网页类型文本与非学术性期刊类型文本、图书类型文本与学术性期刊类型文本之间匹配程度较高，适合用来验证本文提出的特征扩展方法。实验(3)将本文提出的基于《知网》的语义特征扩展方法与实验(2)中未使用特征扩展的分类结果对比，首先比较单一文献类型分类实验扩展前与扩展后的分类效果，分别有了 0.3%、2.3%、1.5%、0.8%，证明本文提出的方法能够提升相同类型文献的分类效果；其次重点比较网页类型文本与非学术性期刊类型文本、图书类型文本与学术性期刊类型文本扩展前与扩展后的分类效果。以网页文献和非学术期刊文献的交叉分类结果为例，扩展后的分类效果分别提升了 1.2%和 2.2%，从而论证了本文提出的基于《知网》的语义特征扩展方法不论对相同类型还是不同类型文献间的自动分类均具有一定的有效性。通过对扩展前后的特征词进行比较，发现在通过特征扩展消除语义差异的过程中，同时也引入了一些“噪声词”，即因在多个类别都频繁出现而类别区分能力低的特征词，从而对识别工作产生干扰，造成分类效果的提高不够显著。另外，虽然网页与非学术期刊的交叉分类效果略低于网页单独分类的实验效果，但是考虑到以非学术期刊作为训练集，可以有效地避免以网页自身作为训练集时需要实时更新所带来的繁冗的工作量，因此文本提出的扩展方法仍然具有实践意义。

5 总结与展望

本文主要研究了借助第三方资源《知网》进行特征扩展的多种类型文献自动分类问题。实验结果证明，在匹配程度较高的文献类型之间，多种类型文献之间的交叉分类可以取得与单一类型文献分类相同甚至更好的分类效果。在此基础上，本文从语料库的构建和特征扩展等两个角度出发，将《知网》作为第三方资源，提出一种基于《知网》的语义特征扩展方法，利用《知网》中的语义结构消除不同文献之间因用词或写作风格等因素造成的差异性问题，以进一步提高多种文献混合自动分类的分类效果，并通过实验证明该方法能有效改进目前分类效果。该方法不仅能够利用经过科学分类、且长期积累的文献信息高效地完成对那些数量增长迅速、更新频繁的文献类型文本的自动分类工作，而且得到了更好的分类效果，因此具有较高的实用性。

本研究是在相对成熟的向量空间模型(VSM)上进行文本表示的基础上完成的。其实质是通过外部资源实现语义扩展，缩小不同类型文献之间的差异，实现跨文献类型分类。今后，在继续深入探究文献类型差异问题的基础上，拟对多种类型文献的混合自动分类中如何消除“噪声词”对分类效果产生的干扰，从而进一步提高分类效果等方面开展更深入的探讨；此外，还可尝试采用概率主题模型(LDA)等文本表示模型对文本建模，以及应用维基百科等其他可能的第三方资源进行跨文献类型分类，评价支持向量机(SVM)等多种经典分类算法对跨文献类型分类的适应性等研究问题。

参考文献：

[1] 薛春香，夏祖奇，侯汉清．基于语料和基于标引经验的自动分类模式比较[J]．南京农业大学学报：社会科学版，2005, 5(4): 85-91. (Xue Chunxiang, Xia Zuqi, Hou Hanqing. A Comparison of Automatic Classification Between Corpus-based Model and Experiences-based Model [J]. Journal of Nanjing Agricultural University: Social Sciences Edition, 2005, 5(4): 85-91.)

[2] Pong J Y H, Kwok R C W, Lau R Y K, et al. A Comparative Study of Two Automatic Document Classification Methods in a Library Setting [J]. Journal of Information Science, 2008, 34(2): 213-230.

[3] 李湘东，胡逸泉，巴志超，等．数字图书馆多种类型文献混合自动分类研究[J]．图书馆杂志，2014, 33(11): 42-48. (Li Xiangdong, Hu Yiquan, Ba Zhichao, et al. The Study of Mixed Automatic Categorization on Digital Library Collections [J]. Library Journal, 2014, 33(11): 42-48.)

[4] 知网[DB/OL]. [2015-06-15]. http://www.keenage.com/. (HowNet Knowledge Database [DB/OL]. [2015-06-15]. http://www.keenage.com/.)

[5] Pan S J, Yang Q. A Survey on Transfer Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.
[6] Wang P, Domeniconi C, Hu J. Using Wikipedia for Co-Clustering Based Cross-Domain Text Classification [C]. In: Proceedings of the 8th IEEE International Conference on Data Mining. IEEE, 2008.
[7] Lu Z, Zhu Y, Pan S J, et al. Source Free Transfer Learning for Text Classification [C]. In: Proceedings of the 28th Association for the Advancement of Artificial Intelligence Conference on Artificia Intelligence. 2014.
[8] 赵辉，刘怀亮．一种基于维基百科的中文短文本分类算法[J]．图书情报工作，2013，57(11): 120-124. (Zhao Hui, Liu Huailiang. Classification Algorithm of Chinese Short Texts Based on Wikipedia [J]. Library and Information Service, 2013, 57(11): 120-124.)
[9] 宁亚辉，樊兴华，吴渝．基于领域词语本体的短文本分类[J]．计算机科学，2009，36(3): 142-145. (Ning Yahui, Fan Xinghua, Wu Yu. Short Text Classification Based on Domain Word Ontology [J]. Computer Science, 2009, 36(3): 142-145.)
[10] 李湘东，曹环，丁丛，等．利用《知网》和领域关键词集扩展方法的短文本分类研究[J]．现代图书情报技术，2015(2): 31-38. (Li Xiangdong, Cao Huan, Ding Cong, et al. Short-text Classification Based on HowNet and Domain Keyword Set Extension [J]. New Technology of Library and Information Service, 2015(2): 31-38.)
[11] 施聪莺，徐朝军，杨晓江．TFIDF 算法研究综述[J]．计算机应用，2009，29(S1): 167-170. (Shi Congying, Xu Chaojun, Yang Xiaojiang. Study of TFIDF Algorithm [J]. Journal of Computer Applications, 2009, 29(S1): 167-170.)
[12] 刘群，李素建．基于《知网》的词汇语义相似度计算[J]．计算语言学及中文语言处理，2002,7(2): 59-76. (Liu Qun, Li Sujian. Word Similarity Computating Based on How-net [J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59-76.)
[13] 吴健，吴朝晖，李莹，等．基于本体论和词汇语义相似度的 Web 服务发现[J]．计算机学报，2005，28(4): 595-602. (Wu Jian, Wu Zhaohui, Li Ying, et al. Web Service Discovery Based on Ontology and Similarity of Words [J]. Chinese Journal of Computers, 2005, 28(4): 595-602.)
[14] 李生琦，田巧燕，汤承．基于《<知网>》词汇语义相关度计算的消歧方法[J]．情报学报，2009，28(5): 706-711．(Li Shengqi, Tian Qiaoyan, Tang Cheng. Disambiguating Method for Computing Relevancy Based on HowNet Semantic Knowledge [J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(5): 706-711.)
[15] 孙建旺，吕学强，张雷瀚．基于语义与最大匹配度的短文本分类研究[J]．计算机工程与设计，2013，34(10): 3613-3618. (Sun Jianwang, Lv Xueqiang, Zhang Leihan. Short Text Classification Based on Semantics and Maximum Matching Degree [J]. Computer Engineering and Design, 2013, 34(10): 3613-3618.)
[16] 搜狗互联网语料库[DB/OL]．[2015-06-03]．http://www.sogou.com/labs/dl/t.html．(SogouT [DB/OL]. [2015-06-03]. http://www.sogou.com/labs/dl/t.html.)
[17] Tan S. An Effective Refinement Strategy for KNN Text Classifier [J]. Expert Systems with Applications, 2006, 30(2): 290-298.
[18] 奉国和．文本分类性能评价研究[J]．情报杂志，2011，30(8): 66-70. (Feng Guohe. Review of Performance Evaluation of Text Classification [J]. Journal of Intelligence, 2011, 30(8): 66-70.)

作者贡献声明:

李湘东：提出研究思路和方案，论文审阅和最终版本修订；
刘康：系统实现，进行实验，论文撰写；
丁丛：进行实验，文献调研；
高凡：数据采集，文献调研。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 李湘东，刘康，丁丛，高凡. book.txt. 图书类型文献.
[2] 李湘东，刘康，丁丛，高凡. web.txt. 网页类型文献.
[3] 李湘东，刘康，丁丛，高凡. acd.txt. 学术性期刊类型文献.
[4] 李湘东，刘康，丁丛，高凡．nonacd.txt．非学术性期刊类型文献.

收稿日期: 2015-08-12
收修改稿日期: 2015-10-19

A New Automatic Categorization Method with Documents Based on HowNet

Li Xiangdong[1,2] Liu Kang[1] Ding Cong[1] Gao Fan[1]
[1](School of Information Management, Wuhan University, Wuhan 430072, China)
[2](Center for the Studies of Information Resources, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This paper aims to solve the feature mismatch problem caused by different document types and improve the performance of automatic classification technology. [**Methods**] We proposed a new method to extend the semantic features using documents of various types as the corpus, which were introduced the third-party resource HowNet and were different with the other un-categorized ones. [**Results**] Compared with the non-feature-extension classification method, the proposed method increased the F-measure by 1.2% to 11.0% in our classification experiment. Four document types, used in our study included webpages, books, non-academic periodicals and academic journals. [**Limitations**] Not every type of document was tested with the publicly accessible corpus, thus, more tests were needed to examine the generalization and objectiveness of the new method. [**Conclusions**] Our study showed that the proposed method was feasible. It could effectively eliminate the semantic differences among various types of collections and improve the performance of automatic text classification through corpus construction and feature extension.

Keywords: Third-party resource HowNet Feature extension Semantic difference

---

爱荷华大学图书馆发布开放获取声明

为了推进爱荷华大学对开放式研究、思想自由和学术作品公共获取的长期承诺, 爱荷华大学图书馆采纳了开放获取政策, 将使其出版物免费获取并确保其长期保存和可发现。该政策完善了图书馆对开放获取的支持, 从而支持自由获取学术作品, 促进员工角色多样化——可充当学术和专业文献的生产者和保存者, 彰显爱荷华大学图书馆宗旨和价值。所有爱荷华大学图书馆的工作人员授予爱荷华大学存储和公开获取他们专业出版物全文的权利。这些出版物包括期刊论文和书籍章节等传统出版物, 并延伸至其他格式的文档, 如会议演示幻灯片、公开演讲的音频和视频记录。该协议为爱荷华大学存储和再发布作品提供了非排他性、全球性、不可撤销的和免版税的权利许可。将尊重出版商的时滞期要求, 在每项作品出版、展示或传播后的三十天内, 向爱荷华大学的机构知识库爱荷华研究在线(Iowa Research Online)提交电子版作品。理想情况下, 提交的版本将是出版商的最终审定稿或作者的最终录用手稿。只有在特殊情况下, 例如一个出版商拒绝接受该政策的条款, 图书馆可发送信息给学术出版团队主席, 选择豁免本政策中的责任。学术出版团队将负责解释该政策、解决相关问题, 并根据需要修改政策。学术出版团队将在采纳该政策一年后对它进行审查, 并将其发现报告给爱荷华大学图书馆。

(编译自: http://www.lib.uiowa.edu/collections/oa-statement/)

(本刊讯)